Vous avez dit TEI ?

Blandine Nouvel

Vous avez dit TEI ?

Blandine NOUVEL
Centre Camille Jullian, UMR 6573 CNRS- Aix-Marseille Université
nouvel@mmsh.univ-aix.fr

De la description des textes numérisés

L'informatique et les sciences humaines ont entamé depuis plusieurs décennies une collaboration fructueuse qui a conduit à l'élaboration de produits éditoriaux numériques originaux. Appliquées aux sciences humaines et sociales, les technologies du numérique transforment les méthodes d'étude des sources de la recherche et l'accès au savoir au travers d'un ensemble de pratiques et de méthodologies regroupées aujourd'hui sous le terme de *Digital Humanities* ou *Humanités numériques*. Parmi les méthodes, les formats et les outils disponibles, la TEI constitue un standard pour la description des données[1].

C'est ainsi que du 9 au 11 juin derniers se sont tenues à Lyon des journées de rencontres et d'échanges sur « La TEI en France, pratiques et perspectives ». Grâce au MutEC[2] soutenu par le TGE ADONIS, une quarantaine de chercheurs en sciences humaines, informaticiens et documentalistes ont trouvé là l'opportunité de présenter des projets aboutis ou en cours sur les sources littéraires et historiques (manuscrites ou imprimées, du Moyen-âge à la Renaissance) ou l'analyse linguistique. Ils se sont confrontés aux arcanes de l'encodage et ont débattu de leurs pratiques et leurs usages de la TEI.

Les secrets du « bon » encodage

Au départ, il y a un texte, élément d'un corpus. S'il est manuscrit, il est transcrit puis numérisé ; imprimé, il sera scanné. Le fichier produit est structuré en XML. Lui seront appliquées des balises de la TEI qui définiront les éléments informationnels et la structure logique du texte originel. Cette opération d'encodage peut parfois être automatisée, du moins en partie, mais le spécialiste devra y mettre nécessairement sa « main » pour vérification et compléments. Le recours à des prestataires extérieurs au groupe projet pourra être profitable, par exemple pour l'acquisition de données ou des développements logiciels.

Exposés théoriques, exercices d'application, partage d'expériences ont donné, chacun à leur façon, les moyens, aux néophytes comme aux plus aguerris, de (tenter) de percer les secrets du « bon » encodage. Une certitude, rien n'est possible sans compréhension du texte, dans sa forme et dans son fond puisque l'encodage doit restituer fidèlement la structure du document et les éléments à décrire. D'où l'importance des métadonnées, définies pour l'ensemble du corpus ou pour une partie (document ou fragment de texte). La TEI n'impose que la description bibliographique comme élément obligatoire dans l'entête du fichier XML. On y définira alors dans un ordre déterminé l'ensemble des métadonnées en prenant soin de veiller à leur compatibilité avec les autres standards. Elles préciseront les éléments relatifs à l'origine du document, à son codage, son profil, ses éventuelles révisions.

D'où aussi le choix épineux du schéma TEI (sélection des balises et spécification de valeurs d'attributs) à définir pour tenir compte au mieux des spécificités de chaque projet. Largement

discuté, documenté, testé et validé au cours de l'avancement du projet, il devra restituer tous les éléments porteurs de sens : la structuration du texte (vers, strophes, parties, paragraphes, segments, notes, signatures, lettrines…) ainsi que les entités non-textuelles (noms propres de lieux ou de personnes, les dates…), en veillant à désambiguïser les termes homonymes. En parallèle, on pourra exploiter un thésaurus, construire des listes d'autorité et inclure des liens vers des éléments différents du corpus (autre document, sous-ensemble) ou associés (images, sons, vidéos). La question de la meilleure méthode reste cependant posée : faut-il intégrer dès le départ toute la bibliothèque TEI pour rejeter finalement les balises inexploitées ou bien sélectionner à priori les balises jugées utiles, quitte à en ajouter plus tard et à devoir corriger l'encodage déjà réalisé ?

D'où enfin, la nécessité de définir les résultats à atteindre : s'agit-il de constituer une publication en ligne, d'organiser et de gérer un corpus pour l'étude, quel niveau de profondeur dans la description choisir ? Rassurons-nous, l'encodage parfait n'existe pas : le compromis se portera sur un rapport temps d'encodage/objectifs et tiendra compte de la capacité des outils et des logiciels de traitement à gérer toute la TEI.

XML-TEI et édition

Le XML structure l'information mais ne constitue pas la forme de lecture idéale! En suivant l'exemple des Bibliothèques Virtuelles Humanistes[3], on peut gérer un site web en XML-TEI via XTF (*XML* Transformation Framework). Plus traditionnellement, on utilisera le langage de transformation XSLT (eXtensible Stylesheet Language Transformations)[4] pour gérer l'affichage des données. Puis un processeur type Xpath[5] appliquera aux nœuds successifs du fichier XML un schéma XSLT défini afin de restituer un document mis en forme. Les applications sont multiples : d'abord rendre lisible un fichier XML pour produire des documents et des outils électroniques de travail répondant directement aux besoins de la recherche, mais aussi élaborer des maquettes sophistiquées, propres à l'édition en ligne ou à la création de supports d'impression.

C'est entre autres parce qu'un seul fichier XML peut fournir autant de supports éditoriaux qu'il y aura de modèles de transformation que le trinôme XML-TEI-XSLT est devenu le cœur de la nouvelle chaîne éditoriale des presses de l'université de Caen. Les PUC ont ainsi su préserver les spécificités des métiers de l'édition et favoriser les relations auteur-éditeur autour du texte, tout en négociant avec succès le tournant du numérique. D'autres expériences convergentes sont conduites[6] mais le modèle doit encore essaimer. Cependant les obstacles demeurent et ne sont pas d'ordre technique, mais plutôt dans les mentalités : la collaboration chercheur-éditeur et la relation à la publication scientifique doivent changer, les versions évolutives d'un texte doivent être intégrées, comme la reconnaissance des publications électroniques dans l'évaluation scientifique.

Une communauté française en devenir

Les projets exposés[7] et ceux recensés en France démontrent une présence forte dans les disciplines des humanités : linguistique, littérature, histoire, dans les domaines de l'archivistique, de l'édition et des bibliothèques. Malgré l'intérêt tout récent du secteur privé aux formations proposées[8], il confirme en France le rôle majeur des structures de l'enseignement supérieur et de la recherche. Les compétences y sont fortes, voire uniques, puisqu'elles y trouvent logiquement la matière même de leur exercice : les corpus sont une tradition humaniste et universitaire.

Reste à organiser la communauté TEI nationale. Le TGE ADONIS est idéalement positionné et sollicité pour jouer un rôle fédérateur. Les utilisateurs et praticiens doivent néanmoins se "prendre en charge", adhérer au consortium, s'entraider et discuter via les outils existants[9]. Dans la mouvance des *Digital Humanities*, une nouvelle compétence émerge, à l'interface de l'informatique, de la recherche et de l'édition.

---

[1] Le consortium Texte Encoding Initiative nait dans les années 1990. Son site fournit l'ensemble des recommandations et outils utiles à l'encodage d'un texte numérique, notamment la dernière version (2007) des *Guidelines for Electronic Text Encoding and Interchange* P5 http://www.tei-c.org/index.xml

[2] MutEC, Mutualisation pour les éditions critiques et les corpus. Associe l'Atelier des Humanités Numériques de l'ENS de Lyon et le Service d'Ingénierie Documentaire de l'Institut des Sciences de l'Homme autour de la diffusion de méthodologies et de technologies pour des projets d'édition critiques et de corpus numériques http://www.mutec-shs.fr/

[3] BVH - Bibliothèques Virtuelles Humanistes du Centre d'Etudes Supérieures de la Renaissance l'université de Tours http://www.bvh.univ-tours.fr/

[4] Le site français sur la XSLT : http://www.xslt.fr/

[5] Xpath : langage de requête pour le XSLT qui permet de sélectionner des parties d'un document en XML

[6] A différents niveaux sont concernées les éditions de l'ENS-LSH Lyon http://www.ens-lyon.eu/, Revues.org, portail des revues en sciences humaines et sociales et plateforme d'édition électronique http://www.revues.org/, l'Association des éditeurs de la recherche et de l'enseignement supérieur qui regroupe 35 éditeurs d'universités, d'écoles d'enseignement supérieur, d'institutions scientifiques http://www.aderes.fr/ et les éditions Quae qui réunissent depuis 2006 les activités éditoriales du Cemagref, du Cirad, de l'Ifremer et de l'Inra http://www.quae.com/

[7] Outre les BVH du CESR, ont été présentés des travaux de l'ATILF, de ICAR, de l'IRHT et de l'Ecole nationale des chartes

[8] A l'Ecole nationale des chartes, au CESR de l'université François-Rabelais à Tours, à l'ENS-LSH de Lyon

[9] Soit la liste de discussion francophone tei-fr@cru.fr, et son wiki https://listes.cru.fr/wiki/tei-fr/index, une traduction partielle des Guidelines réalisée par l'Afnor